# Increasing Object-Level Reconstruction Quality in Single-Image 3D Scene Reconstruction

Anna Ribic  Antonio Oroz  Meikel Kokowski  Franz Srambical
Technical University of Munich
{firstname}.{lastname}@tum.de

## Abstract

## 1. Introduction

While humans can easily infer the 3D structure as well as the complete (panoptic) semantics of a scene from a single image, this task has been a longstanding challenge in the field of computer vision. The task fundamentally prerequisites learning a strong prior of the 3D world. Traditional methods have made significant strides, from generating geometrically coherent structures [9, 31] to learning different instance semantics [13, 20, 29]. More recent approaches directly learn the 3D panoptic semantics as a whole [7, 40], yet they fall short in capturing the intricate details and nuances at the object level. This paper introduces a novel approach to bridge this gap by integrating a specialized object-level model into the reconstruction process, thereby leveraging the specialized model's object-priors.

Our approach models panoptic 3D reconstruction as a two-stage problem. We first use the model of Dahnert et al. [7] to create an initial reconstruction. Then, we leverage the instance masks to extract the object geometries out of the reconstructed scene. We input each of the extracted objects along with cropped images from the scene and text labels into a diffusion model [5] to refine the rough object-level geometries. Finally, we integrate the refined object geometries back into the initial scene reconstruction to obtain a complete and refined panoptic 3D scene reconstruction.

In summary, our main contributions are as follows:

- We propose a novel approach to panoptic 3D reconstruction involving an inference pipeline that leverages object-level reconstruction models to refine the output of a 3D scene reconstruction backbone.
- We qualitatively demonstrate the effectiveness of our approach on the 3D-Front [11] dataset, showing significant improvements over the state-of-the-art.
- We show that fine-tuning SDFusion [5] on the input scene's object distribution (in our case the 3D-Future dataset [12]) significantly improves the quality of the refined objects.
- We propose *weighted masking*, a novel technique to integrate masking uncertainty into the object-level reconstruction process.
- We introduce a conceptually simple yet effective method for shape alignment, which outperforms rigid alignment methods in our experiments.
- We openly release our model code, training and inference pipelines, as well as our newly constructed variation of the 3D-Front dataset to facilitate future research in the field.

## 2. Related Work

**2D panoptic segmentation** 2D panoptic segmentation merges semantic and instance segmentation, providing detailed pixel-level parsing of images, capturing both general categories (semantic segmentation) and individual object identities (instance segmentation) [18]. Since the original task formulation by Kirillov et al. [18], a number of works have been proposed to solve the task [2–4, 19, 21–23, 28, 33, 34, 37–39], while more recent approaches [17] try to unify image segmentation in its entirety.

**Single-view 3D reconstruction** The work by Snavely et al. [32] was the first notable attempt at reconstructing 3D scenes from unordered photo collections. Since then, the field of image-based 3D reconstruction has seen a number of advancements, culminating in the task of single-view 3D reconstruction [6, 9, 16, 26, 29, 31, 35].

**Shape priors** Wu et al. [36] note that the task of single-view 3D reconstruction is non-deterministic, as there are many 3D shapes that can explain a given single-view input, and propose to use shape priors to shape the solution space such that the reconstructed shapes are realistic, but not necessarily the ground truth.

**3D scene understanding** The task of 3D scene understanding and panoptic reconstruction is analogous to its 2D counterpart and aims to infer the 3D structure and semantics of a scene, including the 3D layout, object instances, and their 3D shapes from images [7] or noisy geometry [14, 15]. Dahnert et al. [7] propose a method – henceforth called *Panoptic 3D* – to jointly solve the tasks of 3D scene understanding and single-view 3D reconstruction by lifting features produced by a 2D backbone into a 3D volume of the camera frustrum, and jointly optimizing for geometric reconstruction as well as semantic and instance segmentation.

**Modality-conditioned shape generation** 3D generative models represent objects in a variety of modalities, including point clouds [1, 25], occupancy grids [26], meshes [27], and signed distance functions [30]. Furthermore, these models can also be distinguished by the type of input they take, such as incomplete shapes [8], images [10], text [24, 41], or other modalities [42]. Notably, Cheng et al. [5] propose *SDFusion*, a 3D object reconstruction method conditioned on images, text and geometrical input.

**Datasets** Research in 3D panoptic reconstruction relies heavily on datasets to train and evaluate algorithms. Notable datasets in this domain include ScanNet (cite here) and Replica (cite here), which provide rich annotations for scene understanding tasks.

Among these datasets, the 3D Front dataset stands out for its comprehensive coverage of indoor scenes. Created by Li et al. (cite here), 3D Front offers detailed geometric reconstructions, semantic segmentation, and instance segmentation annotations for various indoor environments, including living rooms, kitchens, and bedrooms. The synthetic 3D dataset contains 6,801 mid-size apartments with 18,797 rooms populated by 3D shapes from the 3D-Future (cite here) dataset. The dataset's high-quality data acquisition process ensures accurate representations, making it a valuable resource for advancing research in 3D panoptic reconstruction.

In an effort to refine the panoptic reconstruction model, we've compiled a custom dataset comprising over 18,000 samples. Leveraging the diverse scenes of the 3D Front dataset, we use BlenderProc (cite here) for randomly sampling camera poses and 2D rendering. Utilizing a C++ pipeline from (cite panoptic reference), we generate annotated 3D geometry within the respective camera frustum

## 3. Method

Given a 2D RGB image, we leverage Panoptic [1] to predict the geometry within the camera frustum of the image, as well as the corresponding 3D semantic- and instance labels. Since panoptic already predicts instances in 2D and uses this as a prior for 3D instance segmentation, we have access to both 2D and 3D instances. We use the 2D instance mask to crop out an RGB image of the detected object and the 3D instance mask to crop out the reconstructed object shape. Then we input the 3D shape, the 2D cropped image and the semantic label (as text) into the SDFusion model.

The SDFusion model uses task specific encoder to encode the conditions (the 2D image and the text) and separately brings the 3D shape into the latent space, in which it applies the latent diffusion process. After applying noise to the latent code, the condition embeddings are concatenated and cross-attended to in the denoising network, in order to modulate the diffusion process. A decoder then reconstructs the shape. The output of SDFusion is front-facing and potentially not matching the orientation of the object in the reconstructed 3D scene. Therefore we employ a registration algorithm to align the reconstructed objects back into the scene.

**Panoptic 3D Scene Reconstruction** Panoptic [1] takes a single RGB image and reconstructs geometry and predicts semantic- and instance segmentation for the corresponding geometry. Panoptic employs a ResNet-18 encoder to compute features of an RGB image. These features are used to predict both a 2D depth map and 2D instance masks using a depth decoder and Mask R-CNN on the ResNet-18 output. The depth map is used to backproject the features into a sparse volumetric grid and the 2D instance mask is propagated to serve as a seed for the 3D instance mask prediction. A U-Net architecture takes the sparse backprojection to predict occupancy, distance field, semantic- and instance labels for each occupancy.

**SDFusion** SDFusion [2] takes a signed distance field as main input and takes an RGB image as well as a text as input to condition the main input on. A variational autoencoder compresses the signed distance field into a latent space representation. Within the latent space, a diffusion process is employed which gradually diffuses the latent code. Before denoising, the conditions are encoded and concatenated to the noisy latent code. Then, a 3D UNet is used to denoise the latent code and an attention mechanism is used to allow the denoising network to attend to the conditions during denoising. The reconstructed latent code is then decoded using the decoder of the variational autoencoder. The result is a reconstructed signed distance field.

**Registration** For the registration we use a method that aligns the reconstructed front-facing SDFusion output. We first align the reconstructed output to the floor in our 3D scene and then rotate in 16 different, uniform positions around its y-axis. We select the orientation which leads to the smallest per-point difference.

## 4. Conclusion

## References


[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2

[2] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J. Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 909–919, 2023. 1

[3] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020.

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 1

[5] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 1, 2

[6] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 1

[7] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34:8282–8293, 2021. 1, 2

[8] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5868–5877, 2017. 2

[9] Maximilian Denninger and Rudolph Triebel. 3d scene reconstruction from a single viewport. In *European Conference on Computer Vision*, pages 51–67. Springer, 2020. 1

[10] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2

[11] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 1

[12] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 1

[13] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019. 1

[14] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019. 2

[15] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 2

[16] Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Perspectivenet: 3d object detection from a single rgb image via perspective points. *Advances in neural information processing systems*, 32, 2019. 1

[17] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2989–2998, 2023. 1

[18] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 1

[19] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J. Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12871–12881, 2022. 1

[20] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 260–277. Springer, 2020. 1

[21] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 1

[22] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 214–223, 2021.

[23] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. 1

[24] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2022. 2

[25] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 2

[26] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1, 2

[27] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019. 2

[28] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129(5):1551–1579, 2021. 1

[29] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1

[30] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[31] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2172–2182, 2019. 1

[32] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 1

[33] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European conference on computer vision*, pages 108–126. Springer, 2020. 1

[34] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5463–5474, 2021. 1

[35] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 1

[36] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–662, 2018. 1

[37] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 1

[38] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2560–2570, 2022.

[39] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *arXiv preprint arXiv:2308.02487*, 2023. 1

[40] Xiang Zhang, Zeyuan Chen, Fangyin Wei, and Zhuowen Tu. Uni-3d: A universal model for panoptic 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2023. 1

[41] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *arXiv preprint arXiv:2306.17115*, 2023. 2

[42] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4176–4186, 2021. 2